Chapter II

LITERATURE REVIEW

2.1 Introduction

THIS chapter reviews the work carried out by the researchers in the area of script recognition. The development and implementation methodologies for handwritten script recognition methods are reviewed in detail which includes, Principal Component Analysis, Neural Network Based Methods, Multi Layer Perceptron, Feed Forward Network, Support Vector Machine, Hidden Markov Model.

From early 1970 researchers taking the efforts in the field of pattern recognition and has had a considerable growth, major application includes character recognition, target detection, biometric identification etc.

2.2 Pre-Processing and Segmentation Techniques

One of the most cited on-line handwritten script recognition technique is the method employed by A. K. Jain et al.[1]. Automatic identification of handwritten script facilities many important applications such as automatic transcription of multilingual documents and search for documents on the web containing a particular script. Increase in usage of handheld devices that accepts handwritten input, has created a growing demand for algorithms that can efficiently analyze and retrieve handwritten data. This paper proposes a method to classify words and lines from an online handwritten document into one of the six major scripts widely used all over the world: Arabic, Devanagari, Cyrillic, Han, Hebrew and Roman. Input page consist of multiscr ipt data containing Cyrillic, Hebrew, Roman, Arabic, Devanagari and Han script. The page is captured by using cross pad. The pen position is sampled at the constant rate of 132 samples per second. During preprocessing the individual strokes are resembled. Gaussian filter is used for smoothing. No restriction was imposed on the style and size of writing. In this work an input data to the script recognizer is a page. Page is segmented into lines, individual lines are segmented into words. Interline distance is calculated to identify the individual lines. Word segmentation is
done using an x axis projection of the text in the line. Segmentation algorithm can not work when different regions of the handwritten text are in different script.

A brief survey of Indian script character recognition is done by U. Pal et. al [2]. A variety of different scripts are used in writing languages throughout the world. In a multiscrypt, multilingual environment, it is essential to know the script used in writing a document before an appropriate character recognition and document analysis algorithm can be chosen. In view of this, several methods for automatic script identification have been developed so far. They mainly belong to two broad categories—structure-based and visual-appearance-based techniques. This survey report gives an overview of the different script identification methodologies under each of these categories. Methods for script identification in online data and video-texts are also presented. It is noted that the research in this field is relatively thin and still more research is to be done, particularly in the case of handwritten documents.

Statistical character structural modeling for Chinese character proposed by In-Jung Kim et.al [3]. Chinese and Japanese languages uses a kanji script. Typical kanji Character is as shown in fig.2.1. Small skew is detected when a document is scanned using optical scanner.

![Typical Kanji Character](image)

**Figure 2.1 Typical Kanji Character**

Character can be analyzed by using two methods, structural and statistical method. Structural method is not popular inspite of advantages. It is having two minor problems, the existing stroke extraction algorithm is not reliable specially for the degraded inputs. Second problem is structural recognizers are not rigorously formulated.

Segmentation of Hindi words had been attempted by M. Hanmandlu et. al [4]. The problem of segmentation is compounded by presence of modifiers. Structural approach is proposed to capture the similarities and differences between structural classes. The segmentation is performed in hierarchical order

1. Separation of upper modifiers and header line from the character.
2. Detecting and segmenting lower modifiers from the character.
3. Detecting whether the character is conjunct or not,
4. Categorization of top modifiers based on Check_Point, Mid_Point and Touching_Point.

The work conclude with the remark, segmentation is a complex task. Structural approach attains the segmentation accuracy around 75%.

Survey For offline recognition of Devanagari Script is presented by R. Jayadevan et.al [5]. From a survey it is noted that, the errors in recognizing printed Devanagari characters is mainly due to incorrect character segmentation of touching or broken characters. Various preprocessing approaches like skew angle and Shirorekha detection in word, noise removal etc. have been surveyed with Different Page segmentation, word segmentation and character segmentation techniques. It concluded with, only a few work has been reported in the area of unconstrained Devanagari handwritten recognition.

A pioneering efforts for the development of handwritten numeral database of Indian scripts have been presented by Ujjawal Bhattacharya et. al [6]. Stated system concerns with the problem of isolated handwritten numeral recognition of widely used Indian scripts the main contribution is, multistage recognition scheme and pioneering development for Indian numerals database. The databases are subjected to the following two distortion transformation,

Rotation - For each training sample image, four others are generated by rotation. Two new samples are created by rotating the sample with random but different angles.
Blurring - All training samples is blurred by applying a Gaussian blurring kernel. Database would be increased by two samples due to this operation. An input image is binarized by Otsu's Thresholding algorithm. The bounding rectangle of an image is also computed. The image is then normalized to a size of 128 x 128. To summaries significant contribution can be listed as real time numeral database creation based on 776 mail pieces collected from

1. Isolated handwritten numerals extracted from the postal code part of their addresses.
2. specially designed form for data collection. The forms are scanned by using 300 dpi flat bed scanner. The preprocessing techniques included are thresholding or recognition in grayscale domain. Three handwritten databases are developed for Devanagari, Bangla and Oriya numerals and stored in TIF format without preprocessing.

The proposed scheme is applicable only for automatic reading of postal code or tabular form documents.

On-Line Chinese character recognition, a typical survey is presented by Cheng- Lin Liu et.al [7]. Chinese characters are used in daily communication by quarter of the word mainly in Asia. Mostly Japanese Kanji and Chinese Kanji characters have mostly similar shape[3]. Some standard database are, TUAT (Tokyo University Of Agriculture and Technology) which is further divided into two parts - The Nakayosi, usually used for classifier learning and Kuchiblue, used for evaluation. Half thinning of a character image and stroke identification is an important part of pre-processing.

Offline handwritten word recognition(HWR) is Proposed by Jaehwa Park [8]. The key ideas are actively and successfully select a subset of features for each word image which provides the minimum required classification accuracy to get a valid answer and to derive a consistent decision metric which works in multi resolution feature space and considers the interrelationship of lexicon at the same time. A recursive architecture based on interaction between flexible character classification and deductive decision making is developed. The recognition process starts from the initial coarse level using a minimum number of features, then increase the discrimination power by adding other features adaptively and recursively until the result is adapted by the decision maker. For the computational aspect of a feasible solution, unified decision metric, recognition confidence, is derived from two measurements: pattern confidence, evaluation of absolute confidence using shape features and lexical confidence evaluation of the relative string dissimilarity in the lexicon He has implemented the same for the US MAIL Address component. Critical Optimization issue is excluded so optimization is the major limitation of this work.
Reena Bajaj et. al [9] describes an efficient and reliable technique for recognition of handwritten numerals. Three different types of features have been used for classification of numerals. In this paper a new approach for recognition of handwritten Devanagari numerals using multiple neural classifiers is used and tried to exploit information about stylistic variations, similarity between numerals and style invariant features. The overview of the scheme presents, the set of first type of features provide coarse shape description of the numerals and are relatively insensitive to minor changes in character shapes. But these features are not expected to remain unchanged for different writing styles. The second class of features consider providing qualitative descriptions of the characters. Theses descriptions encode intrinsic properties of the characters which are expected to be invariant across writing styles and fonts. The third layer in the recognition network is the integration layer. The overall integration architecture which enables fusion of style-specific and style invariant information for character recognition is an important contribution of the mentioned paper. The experimental results show that the approach is Potentially powerful.

Off-Line cursive Handwritten Recognition System is proposed by Andrew Senior et. al [16]. Handwritten Page written by a single author is scanned and send for segmentation. The parameters associated with word and found to be remove are height, slant, slope, stroke width, rotation. Normalization has been incorporated to remove the above parameters. Smoothing and thinning is used to remove the noise from original document, convolution with a 2- dimensional Gaussian filter is used for smoothing and iterative, erosive, thinning algorithm is applied to reduce the strokes.

B. Wegmann et. al [17] focused on the feature specific vector quantization of the images by which compressed image quality is superior. In that the hand written word is a widely varying recognition target. Relatively high computational power and large storage are required to build a successful handwriting recognition system. Since resources are always limited, the tradeoff between the desired recognition rate and required system resources is an optimization problem, and is an important factor to be consider, when handwriting recognition system is designed.
Amit D. et. al [19] states that the Devanagari script poses new challenges due to its highly cursive nature. In proposed algorithm, character is initially subjected to simple noise removal filter. The basic strength of this method is its simplicity since it exploits the inherent cursive nature of script. Based on a reference co-ordinate system, the significant contour of the character are extracted and characterized as a contour set. The input character is scanned from left top to right bottom and then from right bottom to left top and curves extracted constitute the contour set. The curves are further stored in terms of their normalized centroid, normalized lengths and normalized co-ordinates of points that form them. Normalization is performed for character width and height. The proposed approach was experimentally evaluated for range of Devanagari fonts. The first test is aimed at establishing the effectiveness of proposed approach and the test data involved three largely varying fonts i.e. Osho, Surekh and Yogesh. The proposed approach has two major contributions, the first is to represent each character as contour set and use it for matching based on its length, its relative position in reference co-ordinate system and an interpolation scheme. The next involves focusing on the minutiae among similar contours in limited set. This is done by prioritization scheme which concentrates only on those portions of character which reflects its uniqueness. Experimental results constitute significant improvement over contemporary approaches which aim to achieve the same goal.

A modular system to recognize handwritten numerical strings is proposed by Luiz Oliveira et.al. [21]. It uses the segmentation based recognition approach. Main focus has been the recognition of numerical amounts on Brazilian bank checks and numeral strings of NIST SD19(hsf_7 series). Combination of different levels such as segmentation, recognition and post processing is made within multihypothesis approach. Efficient schemes were shown to deal with under segmentation and over segmentation. Comprehensive experiments on numerical amounts NIST SD 19 database have been conducted[110].

Line, word segmentation is proposed by B. B. Chaudhury et. al [43] the work deals with text line identification of Bangla and English, Hindi, Malayalam scripts. Stepwise histogram is drawn to detect local text lines and gaps. Vertical strip width is computed and noise would be detected. Initial curves were drawn by calculating
average standard deviation. Six different scripts text lines were detected with maximum accuracy of 94.31%.

Multi-Script Line identification from Indian documents has been proposed by U. Pal et. al [50]. Automatic scheme is presented to identify text lines of different Indian scripts from a documents. Digitization has been done by flat bed scanner. Histogram based approach is used to convert them into two-tone image. Document skew is successfully detected and corrected. After skew correction the lines are segmented. Horizontal projection profile is used for line segmentation. This work includes the scripts like Bangla, Devanagari, Gurmukhi, English, Malayalam, Tamil, Telugu, Guajarathi, Kannada, Urdu, Oriya. Separation of Bangla, Devanagari and Gurmukhi is tricky because of their structural similarity. Errors are noticed for short line containing fifteen or less characters.

Lines and words are to be identified before applying to the recognition engine. In some documents lines are not parallel to each other, the approach is proposed by U. Pal et. al [69]. Single page may consists of text lines which are not written parallel to each other. i.e. these lines have inclinations with horizontal lines. A robust technique that detects, the arbitrary orientation of the text line, skew angle of individual text is proposed. Bottom-up approach where the components are first labeled and then clustered into the words is proposed. 3045 text lines were used for the experimentation and the work attains an accuracy of 97.7%. Many techniques are proposed to enhance the degraded character image [10], [11], [14], [23], [26], [30], [95], [99], [100].

2.3 Feature Extraction, Classification and Recognition Techniques

Different features are extracted and used for the recognition of Devanagari characters. Sinha et. al [51] for printed Devanagari characters stores structural description for each symbol of the script in terms of primitives and their relationship. Daubechies filter from wavelet domain are considered for multiresolution analysis of handwritten numeral images by U. Bhattacharya et. al [6]. Multistage recognition scheme has been proposed in which the first stage involves a cascade of three Multilayer perceptrons (MLP) classifiers. If a decision about the possible class of an
input numeral image cannot be reached by any of the MLP of the first stage, then conditional probabilities obtained from these classifiers are fed to another MLP in the second stage. The classification accuracy obtained for 20 class handwritten Devanagari - English mixed recognizer is 65.02 percent and 70.85 percent respectively.

Several statistical features are considered by Bansal et. al [12]. Horizontal zero crossing, moments, vertex points and pixel density are the features to be extracted from different zones of Devanagari character. For a word character boxes, number of vertical bars, number of upper modifier boxes, number of lower modifier boxes etc. are considered.

A. W. Senior considered [16] the features like dots, Junctions, endpoints, turning points, and loops. These features are extracted from a single and recorded with line segment features for each horizontal strip. Adding the extra features to the handwriting found to reduce the error. Recurrent error neural networks is used for the recognition with single layer perceptrons and nonlinear activation function. The network is fully connected and trained by using generalized delta rule. To find the best time to stop training, the training set is partitioned into training and validation set. Hidden Markov Model (HMM) is used to derive the best word choice from a sequence of frame probability distributions. Recognition rate achieved is 87% on an open-vocabulary task.

Gradient features have been employed by S. Kompalli et. al [45]. The gradient features are computed by using Sobel operator. It measures the magnitude and direction of intensity changes in a small neighborhood of each pixel.

Sethi and chatterjee [72] described handwritten Devanagari numeral recognition based on structural approach. The primitives used are horizontal and vertical line segments, left and right slants.

Three types of features are used by Reena Bajaj et. al [9] for recognition of Devanagari numerals. The features considered are density features, Moment features of right, left, upper and lower profile curves, Descriptive component features, Neural
Network based classification scheme which consists of multi-classifier connectionist architecture has been used. Multi-classifier consists of
1. Kohonen Self Organizing Map at the lowest layer;
2. A single-layer super-structure laid on Kohonens Net;
3. Multi-layer perceptrons (MLP) based classifier for segment features;

Overall accuracy by using three types of features and multi-classifier connectionist architecture is 89.68% for Devanagari numerals.

The output from different levels such as segmentation, recognition and post processing in probabilistic model has been combined, the approach is proposed by Luiz Oliveira [21] for automatic recognition of handwritten numeral string. NIST SD19 database which is update of SD3 and SD7 is used for experimentation. probabilistic model estimates the most probable interpretation of the written amount for which an input image considered is I. Classifiers used in this system consists of five classes and are shown in fig. 2.1. Multilayer perceptrons classifier (MLP) from neural network is used. Post processing consists of recognition and verification. General-purpose recognizer is used with global decision module (probability module). The recognition rate is up to 93.88 for NIST database and Brazilian bank checks numerical database. The modular system is proven robust for both the database[101]. The work is restricted only for numerals.

Table – 2.1 Five Classes Defined By Oliveira et. al [21]

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Classes Recognized</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_3$</td>
<td>3 non numeral classes { &quot;#&quot;, &quot;,&quot;, &quot;.&quot;}</td>
<td>To feed $e_{13}$</td>
</tr>
<tr>
<td>$e_{10}$</td>
<td>10 Numeral classes { 0-----9 }</td>
<td>To feed $e_{13}$ (numerical amounts) and classification (NIST)</td>
</tr>
<tr>
<td>$e_{13}$</td>
<td>13 classes {0-----9} U { &quot;#&quot;, &quot;,&quot;, &quot;.&quot;}</td>
<td>Classification (numerical amounts)</td>
</tr>
<tr>
<td>$v_0$</td>
<td>Isolated and over segmented characters</td>
<td>Verification</td>
</tr>
<tr>
<td>$v_v$</td>
<td>Isolated and under segmented characters</td>
<td>Verification</td>
</tr>
</tbody>
</table>
M. Blumenstein et. al [55] proposed a feature set based on character pattern. It uses directional, transition features. To prepare a character following Directional feature are used
1. Starting point and intersection point location
2. Distinguish individual line segments
3. Labeling line segment information
4. Line type normalization

Transition features calculates transition from background to foreground pixels in the vertical and horizontal directions. Neural network based classifiers are used which consist of Back propagation and Radial Basis Function Network. The overall accuracy stated is up to 85.48%. Wider variety of local and global features and integration of recognition system is excluded in the present work.

Feature fusion is proposed by Quen -Sen Sun et.al [39]. Extract the two groups of feature vectors with same pattern; establish the correlation criterion function between the two groups. extract canonical correlation features to form effective discriminant vector for recognition. The work has been evaluated by using two database, handwritten Arabic numerals and Yale database for face recognition. Following features are extracted and used for fusion
1. 256-dimensional Gabor transformation feature
2. 21-dimensional Legendre moment feature
3. 36-dimensional Seudo- Zernike moment feature
4. 30-dimensional Zernike moment feature

Minimal distance and Nearest neighborhood are the classifiers used for recognition. The work attains accuracy up to 95.56%.

Veena Bansal et. al [54] performed a preliminary character classification for the printed Devanagari script. These classes are based on the presence of vertical bar
1. NON_BAR characters -
2. END_BAR characters
3. MID_BAR characters

The cardinality of the feature set is obtained after applying horizontal zero crossing varies from 1 to 11. Unknown character is matched with the set obtained after preliminary classification. It attains an accuracy of 90% after post processing.
A contour code feature based segmentation for Roman handwritten script is done by Brijesh Verma [73]. Point of change, Direction change (UP /Down), Number of ascenders, Number of descenders, start point, end point are the features used as a contour code. Neural Network based Back propagation algorithm is used to recognize the segmented characters. Extensive experimentation is conducted on standard CEDAR benchmark dataset. For lexicon size of 20 the classifier attains 91% accuracy.

Comparative study of Devanagari handwritten Character recognition using different features and classifiers is proposed by U. Pal et. al [63]. Four sets of features, two sets from binary and two sets from gray-scale image are used for the recognition process. Computation of Gradient feature, Computation of curvature feature is done. In this work the classifiers considered for the comparisons are Euclidian Distance [ED], Projection Distance [PD], Subspace Method [SM], Modified Quadratic Discriminant Function [MQDF], Modified Projection Distance (MPD), Linear Discriminant Function (LD), Mirror Image Learning (MIL), Support Vector Machine (SVM) etc. From the details of the result it is found that MPD attains 94.15 % the accuracy for single character.

Bounding box is used for the feature extraction and 392 features are obtained for off-line handwritten character recognition of Devanagari Script presented by U. Pal et.al [52]. Recognition of character classifier is carried out by using discriminant function. 5 -fold cross validation scheme is used for recognition. Confusing pair computation is also performed on the like (अ, आ), (ए, ऐ), (ओ, औ) etc. Performance evaluation is done on both Handwritten and printed characters & It attains 80.36% accuracy for handwritten samples and 94.24% for printed samples.

Problem of Feature selection and classification is proposed by Macro Bressan et. al [15]. Classification is done when classes are modeled by statistically independent features. Under the assumption of class-conditional independence the class separability measure of divergence is greatly simplified and becoming sum of undimensional divergence. Divergence and the Bayes decision scheme are adapted to this class- conditional representation. Independent component analysis parameters are estimated. Class - conditional Independent Component analysis is also used. It is
concluded that conditionally independent features greatly simplify pattern classification and feature selection problems.

Self Organizing map has special property of effectively creating spatially organized inter representations of various features of an input signal [24]. Teuvo Kohonen has proposed the classification approach in which competitive learning is used. Two dimensional network of is used. The cells may have hexagonal or rectangular structure. Weight vector is associated with every connection. Based on Euclidean distance between input and weight vector match would be declared. Minimum distance defines the winner. Weights are updated in the learning process. In Self Organizing Map cells or their responses are grouped into subsets corresponding to discrete class of patterns and problem becomes decision process and could be handled separately. Vector Quantization is invited to approximate input signal value or their probability function by quantized "codebook " vectors. Learning vector quantization is an algorithm that defines near optimal decision borders between the classes so used after SOFM.

On line Chinese Character Classification is proposed by Cheng -Lin Liu et. al [7] Classification consists of Coarse classification, Structural matching, Probabilistic matching and statistical classification. The databases considered are TUAT. The accuracy is about 98%. Few of the IEEE and other important papers are surveyed in tabular form. The survey consists of Printed and Handwritten Devanagari Characters Recognition system.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature</th>
<th>Classifier</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kompalli et. al [45 ]</td>
<td>GSC</td>
<td>K- nearest neighbor</td>
<td>1,882</td>
<td>58.51</td>
</tr>
<tr>
<td>Kompalli et. al [46]</td>
<td>GSC</td>
<td>Neural Networks</td>
<td>14,353</td>
<td>61.8</td>
</tr>
<tr>
<td>Kompalli et. al [47 ]</td>
<td>SFSA</td>
<td>Stochastic finite state automation</td>
<td>10,606</td>
<td>87</td>
</tr>
<tr>
<td>Chaudhuri et. al [59]</td>
<td>Statistical</td>
<td>Tree classifier and Template Matching</td>
<td>10,000</td>
<td>83.67</td>
</tr>
<tr>
<td>Huanfeng et. al [74]</td>
<td>Structural Statistical</td>
<td>Hausdroff image comparison</td>
<td>578</td>
<td>66.78</td>
</tr>
<tr>
<td>Govindaraju et.al [75]</td>
<td>Gradient</td>
<td>Neural Networks</td>
<td>4,506</td>
<td>53</td>
</tr>
</tbody>
</table>

Table – 2.2  Details of Printed Devanagari Character Recognition Systems.
### Table 2.3 - Details Of Handwritten Devanagari Word Recognition Systems.

<table>
<thead>
<tr>
<th>Method</th>
<th>Feature</th>
<th>Classifier</th>
<th>Dataset</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Govindaraju et. al[75 ]</td>
<td>Gradient</td>
<td>Neural Networks</td>
<td>4,506</td>
<td>84</td>
</tr>
<tr>
<td>Kompalli et.al [45 ]</td>
<td>GSC</td>
<td>Neural Networks</td>
<td>32,413</td>
<td>84.77</td>
</tr>
<tr>
<td>Bansal et. al [12 ]</td>
<td>Statistical structural</td>
<td>Statistical knowledge Source</td>
<td>Unspecified</td>
<td>87</td>
</tr>
<tr>
<td>Huanfeng et.al [74 ]</td>
<td>Structural</td>
<td>Hausdroff image comparison</td>
<td>2,727</td>
<td>88.24</td>
</tr>
<tr>
<td>Sinha et. al [51 ]</td>
<td>Structural</td>
<td>Syntactic pattern recognition</td>
<td>Unspecified</td>
<td>90</td>
</tr>
<tr>
<td>Bansal et. al [31]</td>
<td>Filters</td>
<td>Five Filter</td>
<td>Unspecified</td>
<td>93</td>
</tr>
<tr>
<td>Dhurandhar et. al [19]</td>
<td>Contours</td>
<td>Interpolation</td>
<td>546</td>
<td>93.03</td>
</tr>
<tr>
<td>Kompalli et. al [46]</td>
<td>GSC</td>
<td>K-nearest neighbor</td>
<td>9,297</td>
<td>95</td>
</tr>
<tr>
<td>Chaudhury et. al [43 ]</td>
<td>Statistical</td>
<td>Tree Classifier and Template Matching</td>
<td>10,000</td>
<td>95.08</td>
</tr>
<tr>
<td>Kompalli et. al [47]</td>
<td>SFSA</td>
<td>Stochastic Finite State Automation</td>
<td>10,606</td>
<td>96</td>
</tr>
</tbody>
</table>

### 2.4 Concluding Remarks

Literature survey reveals that benchmarking database for English, Chinese and Japanese language are available (i.e. for Roman and Kanji script). There is no benchmarking database for Devanagari script. Methods proposed by researchers are mostly for recognition of Devanagari numerals & characters. Comparatively less work has been reported for offline handwritten Devanagari script recognition. Line and word segmentation is complex and intricate to implement. Cognizance of survey also reveals that less work has been reported for compound character recognition. In this research the following algorithms are developed.
1. Page, Line, Character segmentation.
2. Recognition of compound characters.
3. Devanagari script recognition based on Vector Quantization.
4. Recognition results are further compared with Support Vector.